DATA file
LPGA2014

statistics for 134 golfers for 2014 appear in the file named *LPGA2014* (LPGA website, April 2015). Earnings ($1000s) is the total earnings in thousands of dollars; Scoring Avg. is the scoring average for all events; Greens in Reg. is the percentage of time a player is able to hit the greens in regulation; and Putting Avg. is the average number of putts taken on greens hit in regulation. A green is considered hit in regulation if any part of the ball is touching the putting surface and the difference between par for the hole and the number of strokes taken to hit the green is at least 2.

a. Develop an estimated regression equation that can be used to predict the scoring average given the percentage of time a player is able to hit the greens in regulation and the average number of putts taken on green hit in regulation.
b. Plot the standardized residuals against $\hat{y}$. Does the residual plot support the assumption about ϵ? Explain.
c. Check for any outliers. What are your conclusions?
d. Are there any influential observations? Explain.

15.9 Logistic Regression

In many regression applications, the dependent variable may only assume two discrete values. For instance, a bank might want to develop an estimated regression equation for predicting whether a person will be approved for a credit card. The dependent variable can be coded as $y = 1$ if the bank approves the request for a credit card and $y = 0$ if the bank rejects the request for a credit card. Using logistic regression we can estimate the probability that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

Let us consider an application of logistic regression involving a direct mail promotion being used by Simmons Stores. Simmons owns and operates a national chain of women's apparel stores. Five thousand copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a $50 discount on purchases of $200 or more. The catalogs are expensive and Simmons would like to send them to only those customers who have a high probability of using the coupon.

Management believes that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon. Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card. Simmons sent the catalog to each of the 100 customers selected. At the end of a test period, Simmons noted whether each customer had used her or his coupon. The sample data for the first 10 catalog recipients are shown in Table 15.11.

DATA file
Simmons

TABLE 15.11 Partial Sample Data for the Simmons Stores Example

Customer	Annual Spending ($1000)	Simmons Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not. In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

We might think of building a multiple regression model using the data in Table 15.11 to help Simmons estimate whether a catalog recipient will use the coupon. We would use Annual Spending ($1000) and Simmons Card as independent variables and Coupon as the dependent variable. Because the dependent variable may only assume the values of 0 or 1, however, the ordinary multiple regression model is not applicable. This example shows the type of situation for which logistic regression was developed. Let us see how logistic regression can be used to help Simmons estimate which type of customer is most likely to take advantage of their promotion.

Logistic Regression Equation

In many ways logistic regression is like ordinary regression. It requires a dependent variable, y, and one or more independent variables. In multiple regression analysis, the mean or expected value of y is referred to as the multiple regression equation.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.26}$$

In logistic regression, statistical theory as well as practice has shown that the relationship between $E(y)$ and $x_1, x_2, \ldots, x_p$ is better described by the following nonlinear equation.

LOGISTIC REGRESSION EQUATION

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \tag{15.27}$$

If the two values of the dependent variable y are coded as 0 or 1, the value of $E(y)$ in equation (15.27) provides the *probability* that $y = 1$ given a particular set of values for the independent variables $x_1, x_2, \ldots, x_p$. Because of the interpretation of $E(y)$ as a probability, the **logistic regression equation** is often written as follows:

INTERPRETATION OF *E(y)* AS A PROBABILITY IN LOGISTIC REGRESSION

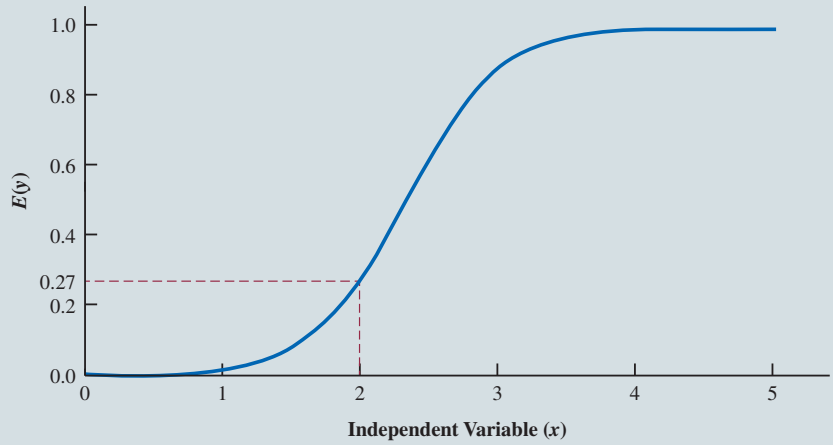
$$E(y) = P(y = 1 | x_1, x_2, \ldots, x_p) \tag{15.28}$$

To provide a better understanding of the characteristics of the logistic regression equation, suppose the model involves only one independent variable x and the values of the model parameters are $\beta_0 = -7$ and $\beta_1 = 3$. The logistic regression equation corresponding to these parameter values is

$$E(y) = P(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}} \tag{15.29}$$

Figure 15.12 shows a graph of equation (15.29). Note that the graph is S-shaped. The value of $E(y)$ ranges from 0 to 1. For example, when $x = 2$, $E(y)$ is approximately .27. Also note that the value of $E(y)$ gradually approaches 1 as the value of x becomes larger and the value of $E(y)$ approaches 0 as the value of x becomes smaller. For example, when $x = 2$, $E(y) = .269$. Note also that the values of $E(y)$, representing probability, increase fairly rapidly as x increases from 2 to 3. The fact that the values of $E(y)$ range from 0 to 1 and that the curve is S-shaped makes equation (15.29) ideally suited to model the probability the dependent variable is equal to 1.

FIGURE 15.12 Logistic Regression Equation for $\beta_0 = -7$ and $\beta_1 = 3$



Estimating the Logistic Regression Equation

In simple linear and multiple regression the least squares method is used to compute b_0, $b_1, \ldots, b_p$ as estimates of the model parameters $(\beta_0, \beta_1, \ldots, \beta_p)$. The nonlinear form of the logistic regression equation makes the method of computing estimates more complex and beyond the scope of this text. We use statistical software to provide the estimates. The **estimated logistic regression equation** is

ESTIMATED LOGISTIC REGRESSION EQUATION

$$\hat{y} = \text{estimate of } P(y = 1 | x_1, x_2, \ldots, x_p) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} \quad \textbf{(15.30)}$$

Here, $\hat{y}$ provides an estimate of the probability that $y = 1$ given a particular set of values for the independent variables.

Let us now return to the Simmons Stores example. The variables in the study are defined as follows:

$$y = \begin{cases} 0 \text{ if the customer did not use the coupon} \\ 1 \text{ if the customer used the coupon} \end{cases}$$

$x_1 =$ annual spending at Simmons Stores ($1000s)

$$x_2 = \begin{cases} 0 \text{ if the customer does not have a Simmons credit card} \\ 1 \text{ if the customer has a Simmons credit card} \end{cases}$$

Thus, we choose a logistic regression equation with two independent variables.

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}} \quad \textbf{(15.31)}$$

In Appendix 15.2 we show how JMP is used to generate the output in Figure 15.13.

Using the sample data (see Table 15.11), we used statistical software to compute estimates of the model parameters β_0, β_1, and β_2. Figure 15.13 displays output commonly

provided by statistical software. We see that $b_0 = -2.146$, $b_1 = .342$, and $b_2 = 1.099$. Thus, the estimated logistic regression equation is

$$\hat{y} = \frac{e^{b_0+b_1x_1+b_2x_2}}{1 + e^{b_0+b_1x_1+b_2x_2}} = \frac{e^{-2.146+.342x_1+1.099x_2}}{1 + e^{-2.146+.342x_1+1.099x_2}} \quad \textbf{(15.32)}$$

We can now use equation (15.32) to estimate the probability of using the coupon for a particular type of customer. For example, to estimate the probability of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card, we substitute $x_1 = 2$ and $x_2 = 0$ into equation (15.32).

$$\hat{y} = \frac{e^{-2.146+.342(2)+1.099(0)}}{1 + e^{-2.146+.342(2)+1.099(0)}} = \frac{e^{-1.462}}{1 + e^{-1.462}} = \frac{.2318}{1.2318} = .1882$$

Thus, an estimate of the probability of using the coupon for this particular group of customers is approximately .19. Similarly, to estimate the probability of using the coupon for customers who spent \$2000 last year and have a Simmons credit card, we substitute $x_1 = 2$ and $x_2 = 1$ into equation (15.32).

$$\hat{y} = \frac{e^{-2.146+.342(2)+1.099(1)}}{1 + e^{-2.146+.342(2)+1.099(1)}} = \frac{e^{-.363}}{1 + e^{-.363}} = \frac{.6956}{1.6956} = .4102$$

Thus, for this group of customers, the probability of using the coupon is approximately .41. It appears that the probability of using the coupon is much higher for customers with a Simmons credit card. Before reaching any conclusions, however, we need to assess the statistical significance of our model.

Testing for Significance

Testing for significance in logistic regression is similar to testing for significance in multiple regression. First we conduct a test for overall significance. For the Simmons Stores example, the hypotheses for the test of overall significance follow:

H_0: $\beta_1 = \beta_2 = 0$
H_a: One or both of the parameters is not equal to zero

The test for overall significance is based upon the value of a χ^2 test statistic. If the null hypothesis is true, the sampling distribution of χ^2 follows a chi-square distribution with degrees of freedom equal to the number of independent variables in the model. While the calculations behind the computation of χ^2 is beyond the scope of the book, Figure 15.13 lists the value of χ^2 and its corresponding p-value in the Whole Model row of the Significance Tests table; we see that the value of χ^2 is 13.63, its degrees of freedom are 2, and its p-value is .0011. Thus, at any level of significance $\alpha \geq .0011$, we would reject the null hypothesis and conclude that the overall model is significant.

If the χ^2 test shows an overall significance, another χ^2 test can be used to determine whether each of the individual independent variables is making a significant contribution to the overall model. For the independent variables x_i, the hypotheses are

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

The test of significance for an independent variable is also based upon the value of a χ^2 test statistic. If the null hypothesis is true, the sampling distribution of χ^2 follows a chi-square distribution with one degree of freedom. The Spending and Card rows of the Significance Tests table of Figure 15.13 contain the values of χ^2 and their corresponding p-values test for the estimated coefficients. Suppose we use $\alpha = .05$ to test for the significance of the

FIGURE 15.13 Logistic Regression Output for the Simmons Stores Example

Significance Tests

Term	Degrees of Freedom	χ^2	p-Value
Whole Model	2	13.63	.0011
Spending	1	7.56	.0060
Card	1	6.41	.0013

Parameter Estimates

Term	Estimate	Standard Error
Intercept	−2.146	.577
Spending	.342	.129
Card	1.099	.44

Odds Ratios

Term	Odds Ratio	Lower 95%	Upper 95%
Spending	1.4073	1.0936	1.8109
Card	3.0000	1.2550	7.1730

independent variables in the Simmons model. For the independent variable Spending (x_1) the χ^2 value is 7.56 and the corresponding p-value is .0060. Thus, at the .05 level of significance we can reject H_0: $\beta_1 = 0$. In a similar fashion we can also reject H_0: $\beta_2 = 0$ because the p-value corresponding to Card's $\chi^2 = 6.41$ is .0013. Hence, at the .05 level of significance, both independent variables are statistically significant.

Managerial Use

We described how to develop the estimated logistic regression equation and how to test it for significance. Let us now use it to make a decision recommendation concerning the Simmons Stores catalog promotion. For Simmons Stores, we already computed $P(y = 1|x_1 = 2, x_2 = 1) = .4102$ and $P(y = 1|x_1 = 2, x_2 = 0) = .1881$. These probabilities indicate that for customers with annual spending of $2000 the presence of a Simmons credit card increases the probability of using the coupon. In Table 15.12 we show estimated probabilities for values of annual spending ranging from $1000 to $7000 for both customers who have a Simmons credit card and customers who do not have a Simmons credit card. How can Simmons use this information to better target customers for the new promotion? Suppose Simmons wants to send the promotional catalog only to customers who have a .40 or higher probability of using the coupon. Using the estimated probabilities in Table 15.12, Simmons promotion strategy would be:

Customers who have a Simmons credit card: Send the catalog to every customer who spent $2000 or more last year.

Customers who do not have a Simmons credit card: Send the catalog to every customer who spent $6000 or more last year.

Looking at the estimated probabilities further, we see that the probability of using the coupon for customers who do not have a Simmons credit card but spend $5000 annually is .3922. Thus, Simmons may want to consider revising this strategy by including

TABLE 15.12 Estimated Probabilities for Simmons Stores

		Annual Spending						
		$1000	$2000	$3000	$4000	$5000	$6000	$7000
Credit Card	Yes	.3307	.4102	.4948	.5796	.6599	.7320	.7936
	No	.1414	.1881	.2460	.3148	.3927	.4765	.5617

those customers who do not have a credit card, as long as they spent $5000 or more last year.

Interpreting the Logistic Regression Equation

Interpreting a regression equation involves relating the independent variables to the business question that the equation was developed to answer. With logistic regression, it is difficult to interpret the relation between the independent variables and the probability that $y = 1$ directly because the logistic regression equation is nonlinear. However, statisticians have shown that the relationship can be interpreted indirectly using a concept called the odds ratio.

The **odds in favor of an event occurring** is defined as the probability the event will occur divided by the probability the event will not occur. In logistic regression the event of interest is always $y = 1$. Given a particular set of values for the independent variables, the odds in favor of $y = 1$ can be calculated as follows:

$$\text{odds} = \frac{P(y = 1|x_1, x_2, \ldots, x_p)}{P(y = 0|x_1, x_2, \ldots, x_p)} = \frac{P(y = 1|x_1, x_2, \ldots, x_p)}{1 - P(y = 1|x_1, x_2, \ldots, x_p)} \quad \textbf{(15.33)}$$

The **odds ratio** measures the impact on the odds of a one-unit increase in only one of the independent variables. The odds ratio is the odds that $y = 1$ given that one of the independent variables has been increased by one unit (odds_1) divided by the odds that $y = 1$ given no change in the values for the independent variables (odds_0).

ODDS RATIO

$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0} \quad \textbf{(15.34)}$$

For example, suppose we want to compare the odds of using the coupon for customers who spend $2000 annually and have a Simmons credit card ($x_1 = 2$ and $x_2 = 1$) to the odds of using the coupon for customers who spend $2000 annually and do not have a Simmons credit card ($x_1 = 2$ and $x_2 = 0$). We are interested in interpreting the effect of a one-unit increase in the independent variable x_2. In this case

$$\text{odds}_1 = \frac{P(y = 1|x_1 = 2, x_2 = 1)}{1 - P(y = 1|x_1 = 2, x_2 = 1)}$$

and

$$\text{odds}_0 = \frac{P(y = 1|x_1 = 2, x_2 = 0)}{1 - P(y = 1|x_1 = 2, x_2 = 0)}$$

Previously we showed that an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 1$ is .4102, and an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 0$ is .1881. Thus,

$$\text{estimate of odds}_1 = \frac{.4102}{1 - .4102} = .6956$$

and

$$\text{estimate of odds}_0 = \frac{.1881}{1 - .1881} = .2318$$

The estimated odds ratio is

$$\text{estimated odds ratio} = \frac{.6956}{.2318} = 3.00$$

Thus, we can conclude that the estimated odds in favor of using the coupon for customers who spent \$2000 last year and have a Simmons credit card are 3 times greater than the estimated odds in favor of using the coupon for customers who spent \$2000 last year and do not have a Simmons credit card.

The odds ratio for each independent variable is computed while holding all the other independent variables constant. But it does not matter what constant values are used for the other independent variables. For instance, if we computed the odds ratio for the Simmons credit card variable (x_2) using \$3000, instead of \$2000, as the value for the annual spending variable (x_1), we would still obtain the same value for the estimated odds ratio (3.00). Thus, we can conclude that the estimated odds of using the coupon for customers who have a Simmons credit card are 3 times greater than the estimated odds of using the coupon for customers who do not have a Simmons credit card.

The odds ratio is standard output for most statistical software packages. The Odds Ratios table in Figure 15.13 contains the estimated odds ratios for each of the independent variables. The estimated odds ratio for Spending (x_1) is 1.4073 and the estimated odds ratio for Card (x_2) is 3.0000. We already showed how to interpret the estimated odds ratio for the binary independent variable x_2. Let us now consider the interpretation of the estimated odds ratio for the continuous independent variable x_1.

The value of 1.4073 in the Odds Ratio column of the output tells us that the estimated odds in favor of using the coupon for customers who spent \$3000 last year is 1.4073 times greater than the estimated odds in favor of using the coupon for customers who spent \$2000 last year. Moreover, this interpretation is true for any one-unit change in x_1. For instance, the estimated odds in favor of using the coupon for someone who spent \$5000 last year is 1.4073 times greater than the odds in favor of using the coupon for a customer who spent \$4000 last year. But suppose we are interested in the change in the odds for an increase of more than one unit for an independent variable. Note that x_1 can range from 1 to 7. The odds ratio given by the output does not answer this question. To answer this question we must explore the relationship between the odds ratio and the regression coefficients.

A unique relationship exists between the odds ratio for a variable and its corresponding regression coefficient. For each independent variable in a logistic regression equation it can be shown that

$$\text{Odds ratio} = e^{\beta_i}$$

To illustrate this relationship, consider the independent variable x_1 in the Simmons example. The estimated odds ratio for x_1 is

$$\text{Estimated odds ratio} = e^{b_1} = e^{.342} = 1.407$$

Similarly, the estimated odds ratio for x_2 is

$$\text{Estimated odds ratio} = e^{b_2} = e^{1.099} = 3.000$$

This relationship between the odds ratio and the coefficients of the independent variables makes it easy to compute estimated odds ratios once we develop estimates of the model

parameters. Moreover, it also provides us with the ability to investigate changes in the odds ratio of more than or less than one unit for a continuous independent variable.

The odds ratio for an independent variable represents the change in the odds for a one-unit change in the independent variable holding all the other independent variables constant. Suppose that we want to consider the effect of a change of more than one unit, say c units. For instance, suppose in the Simmons example that we want to compare the odds of using the coupon for customers who spend \$5000 annually ($x_1 = 5$) to the odds of using the coupon for customers who spend \$2000 annually ($x_1 = 2$). In this case $c = 5 - 2 = 3$ and the corresponding estimated odds ratio is

$$e^{cb_1} = e^{3(.342)} = e^{1.026} = 2.79$$

This result indicates that the estimated odds of using the coupon for customers who spend \$5000 annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend \$2000 annually. In other words, the estimated odds ratio for an increase of \$3000 in annual spending is 2.79.

In general, the odds ratio enables us to compare the odds for two different events. If the value of the odds ratio is 1, the odds for both events are the same. Thus, if the independent variable we are considering (such as Simmons credit card status) has a positive impact on the probability of the event occurring, the corresponding odds ratio will be greater than 1. Most statistical software packages provide a confidence interval for the odds ratio. The Odds Ratio table in Figure 15.13 provides a 95% confidence interval for each of the odds ratios. For example, the point estimate of the odds ratio for x_1 is 1.4073 and the 95% confidence interval is 1.0936 to 1.8109. Because the confidence interval does not contain the value of 1, we can conclude that x_1 has a significant relationship with the estimated odds ratio. Similarly, the 95% confidence interval for the odds ratio for x_2 is 1.2550 to 7.1730. Because this interval does not contain the value of 1, we can also conclude that x_2 has a significant relationship with the odds ratio.

Logit Transformation

An interesting relationship can be observed between the odds in favor of $y = 1$ and the exponent for e in the logistic regression equation. It can be shown that

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

This equation shows that the natural logarithm of the odds in favor of $y = 1$ is a linear function of the independent variables. This linear function is called the **logit**. We will use the notation $g(x_1, x_2, \ldots, x_p)$ to denote the logit.

LOGIT

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.35)}$$

Substituting $g(x_1, x_2, \ldots, x_p)$ for $\beta_1 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ in equation (15.27), we can write the logistic regression equation as

$$E(y) = \frac{e^{g(x_1, x_2, \ldots, x_p)}}{1 + e^{g(x_1, x_2, \ldots, x_p)}} \qquad \textbf{(15.36)}$$

Once we estimate the parameters in the logistic regression equation, we can compute an estimate of the logit. Using $\hat{g}(x_1, x_2, \ldots, x_p)$ to denote the **estimated logit**, we obtain

ESTIMATED LOGIT

$$\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad \textbf{(15.37)}$$

Thus, in terms of the estimated logit, the estimated regression equation is

$$\hat{y} = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}} = \frac{e^{\hat{g}(x_1,x_2,\ldots,x_p)}}{1+e^{\hat{g}(x_1,x_2,\ldots,x_p)}} \tag{15.38}$$

For the Simmons Stores example, the estimated logit is

$$\hat{g}(x_1, x_2) = -2.146 + .342x_1 + 1.099x_2$$

and the estimated regression equation is

$$\hat{y} = \frac{e^{\hat{g}(x_1,x_2)}}{1+e^{\hat{g}(x_1,x_2)}} = \frac{e^{-2.146+.342x_1+1.099x_2}}{1+e^{-2.146+.342x_1+1.099x_2}}$$

Thus, because of the unique relationship between the estimated logit and the estimated logistic regression equation, we can compute the estimated probabilities for Simmons Stores by dividing $e^{\hat{g}(x_1,x_2)}$ by $1 + e^{\hat{g}(x_1,x_2)}$.

NOTES + COMMENTS

1. Because of the unique relationship between the estimated coefficients in the model and the corresponding odds ratios, the overall test for significance based upon the χ^2 statistic is also a test of overall significance for the odds ratios. In addition, the χ^2 test for the individual significance of a model parameter also provides a statistical test of significance for the corresponding odds ratio.
2. In simple and multiple regression, the coefficient of determination is used to measure the goodness of fit. In logistic regression, no single measure provides a similar interpretation. A discussion of goodness of fit is beyond the scope of our introductory treatment of logistic regression.

EXERCISES

Applications

DATA*file* Simmons

44. **Coupon Redemption.** Refer to the Simmons Stores example introduced in this section. The dependent variable is coded as $y = 1$ if the customer used the coupon and 0 if not. Suppose that the only information available to help predict whether the customer will use the coupon is the customer's credit card status, coded as $x = 1$ if the customer has a Simmons credit card and $x = 0$ if not.
    a. Write the logistic regression equation relating x to y.
    b. What is the interpretation of $E(y)$ when $x = 0$?
    c. For the Simmons data in Table 15.11, use statistical software to compute the estimated logit.
    d. Use the estimated logit computed in part (c) to estimate the probability of using the coupon for customers who do not have a Simmons credit card and to estimate the probability of using the coupon for customers who have a Simmons credit card.
    e. What is the estimated odds ratio? What is its interpretation?
45. **Odds Ratio for Coupon Redemption.** In Table 15.12 we provided estimates of the probability of using the coupon in the Simmons Stores catalog promotion. A different value is obtained for each combination of values for the independent variables.
    a. Compute the odds in favor of using the coupon for a customer with annual spending of \$4000 who does not have a Simmons credit card ($x_1 = 4$, $x_2 = 0$).
    b. Use the information in Table 15.12 and part (a) to compute the odds ratio for the Simmons credit card variable $x_2 = 0$, holding annual spending constant at $x_1 = 4$.

c. In the text, the odds ratio for the credit card variable was computed using the information in the $2000 column of Table 15.12. Did you get the same value for the odds ratio in part (b)?

46. **Direct Deposit.** Community Bank would like to increase the number of customers who use payroll direct deposit. Management is considering a new sales campaign that will require each branch manager to call each customer who does not currently use payroll direct deposit. As an incentive to sign up for payroll direct deposit, each customer contacted will be offered free checking for two years. Because of the time and cost associated with the new campaign, management would like to focus their efforts on customers who have the highest probability of signing up for payroll direct deposit. Management believes that the average monthly balance in a customer's checking account may be a useful predictor of whether the customer will sign up for direct payroll deposit. To investigate the relationship between these two variables, Community Bank tried the new campaign using a sample of 50 checking account customers who do not currently use payroll direct deposit. The sample data show the average monthly checking account balance (in hundreds of dollars) and whether the customer contacted signed up for payroll direct deposit (coded 1 if the customer signed up for payroll direct deposit and 0 if not). The data are contained in the data set named Bank; a portion of the data follows.

DATA*file* Bank

Customer	x = Monthly Balance	y = Direct Deposit
1	1.22	0
2	1.56	0
3	2.10	0
4	2.25	0
5	2.89	0
6	3.55	0
7	3.56	0
8	3.65	1
.	.	.
.	.	.
.	.	.
48	18.45	1
49	24.98	0
50	26.05	1

a. Write the logistic regression equation relating x to y.
b. For the Community Bank data, use statistical software to compute the estimated logistic regression equation.
c. Conduct a test of significance using the χ^2 test statistic. Use $\alpha = .05$.
d. Estimate the probability that customers with an average monthly balance of $1000 will sign up for direct payroll deposit.
e. Suppose Community Bank only wants to contact customers who have a .50 or higher probability of signing up for direct payroll deposit. What is the average monthly balance required to achieve this level of probability?
f. What is the estimated odds ratio? What is its interpretation?

47. **College Retention.** Over the past few years the percentage of students who leave Lakeland College at the end of the first year has increased. Last year Lakeland started a voluntary one-week orientation program to help first-year students adjust to campus life. If Lakeland is able to show that the orientation program has a positive effect on retention, they will consider making the program a requirement for all first-year students. Lakeland's administration also suspects that students with lower GPAs have a higher probability of leaving Lakeland at the end of the first year. In order to investigate the relation of

these variables to retention, Lakeland selected a random sample of 100 students from last year's entering class. The data are contained in the data set named Lakeland; a portion of the data follows.

DATA*file* Lakeland

Student	GPA	Program	Return
1	3.78	1	1
2	2.38	0	1
3	1.30	0	0
4	2.19	1	0
5	3.22	1	1
6	2.68	1	1
.	.	.	.
.	.	.	.
.	.	.	.
98	2.57	1	1
99	1.70	1	1
100	3.85	1	1

The dependent variable was coded as $y = 1$ if the student returned to Lakeland for the sophomore year and $y = 0$ if not. The two independent variables are:

$x_1 =$ GPA at the end of the first semester

$$x_2 = \begin{cases} 0 \text{ if the student did not attend the orientation program} \\ 1 \text{ if the student attended the orientation program} \end{cases}$$

a. Write the logistic regression equation relating x_1 and x_2 to y.
b. What is the interpretation of $E(y)$ when $x_2 = 0$?
c. Use both independent variables and statistical software to compute the estimated logit.
d. Conduct a test for overall significance using $\alpha = .05$.
e. Use $\alpha = .05$ to determine whether each of the independent variables is significant.
f. Use the estimated logit computed in part (c) to estimate the probability that students with a 2.5 grade point average who did not attend the orientation program will return to Lakeland for their sophomore year. What is the estimated probability for students with a 2.5 grade point average who attended the orientation program?
g. What is the estimated odds ratio for the orientation program? Interpret it.
h. Would you recommend making the orientation program a required activity? Why or why not?

48. **Repeat Sales.** The Tire Rack maintains an independent consumer survey to help drivers help each other by sharing their long-term tire experiences. The data contained in the file named TireRatings show survey results for 68 all-season tires. Performance traits are rated using the following 10-point scale.

DATA*file* TireRatings

Superior		Excellent		Good		Fair		Unacceptable	
10	9	8	7	6	5	4	3	2	1

The values for the variable labeled Wet are the average of the ratings for each tire's wet traction performance and the values for the variable labeled Noise are the average of the ratings for the noise level generated by each tire. Respondents were also asked whether they would buy the tire again using the following 10-point scale:

Definitely		Probably		Possibly		Probably Not		Definitely Not	
10	9	8	7	6	5	4	3	2	1

The values for the variable labeled Buy Again are the average of the buy-again responses. For the purposes of this exercise, we created the following binary dependent variable:

$$\text{Purchase} = \begin{cases} 1 \text{ if the value of the Buy-Again variable is 7 or greater} \\ 0 \text{ if the value of the Buy-Again variable is less than 7} \end{cases}$$

Thus, if Purchase = 1, the respondent would probably or definitely buy the tire again.

a. Write the logistic regression equation relating x_1 = Wet performance rating and x_2 = Noise performance rating to y = Purchase.
b. Use statistical software to compute the estimated logit.
c. Use the estimated logit to compute an estimate of the probability that a customer will probably or definitely purchase a particular tire again with a Wet performance rating of 8 and a Noise performance rating of 8.
d. Suppose that the Wet and Noise performance ratings were 7. How does that affect the probability that a customer will probably or definitely purchase a particular tire again with these performance ratings?
e. If you were the CEO of a tire company, what do the results for parts (c) and (d) tell you?

15.10 Practical Advice: Big Data and Hypothesis Testing in Multiple Regression

In Chapter 14, we observed that in simple linear regression, the p-value for the test of the hypothesis H_0: $\beta_1 = 0$ decreases as the sample size increases. Likewise, for a given level of confidence, the confidence interval for β_1, the confidence interval for the mean value of y, and the prediction interval for an individual value of y each narrows as the sample size increases. These results extend to multiple regression. As the sample size increases:

- the p-value for the F test used to determine whether a significant relationship exists between the dependent variable and the set of all independent variables in the regression model decreases;
- the p-value for each of t-test used to determine whether a significant relationship exists between the dependent variable and an individual independent variable in the regression model decreases;
- the confidence interval for the slope parameter associated with each individaul independent variable narrow;
- the confidence interval for the mean value of y narrows;
- the prediction interval for an individual value of y narrows.

Thus the interval estimates for the slope parameter associated with each individual independent variable, the mean value of y, and predicted individual value of y will become more precise as the sample size increases. And we are more likely to reject the hypothesis that a relationship does not exist between the dependent variable and the set of all individual independent variable in the model as the sample size increases. And for each individual independent variable, we are more likely to reject the hypothesis that a relationship does not exist between the dependent variable and the individual independent variable as the sample size increases. Even when servere multicollinearity is present, if the sample is sufficiently large, independent variable that are highly correlated may each have a significant relationship with the dependent variable. But this does not necessariloy mean that these results become more reliable as the sample size increases.

No matter how large the sample used to estimate the multiple regression model, we must be concerned about the potential presence of nonsampling error in the data. It is important to carefully consider whether a random sample of the population of interest has actually been taken. If nonsampling error is introduced in the data cololection process, the